Research Strategy

A Background & Significance

Human diseases are influenced by many genetic variants that often act in concert to effect change in cellular function [12]. The overwhelming majority of these mutations lies in non-coding regions and are enriched within regulatory elements [1, 38, 9]. In particular, they are over-represented in one particular class of variant, expression quantitative trait loci (eQTL), that associate the presence of a genetic variant with the expression level of a gene [22, 7, 30, 23]. They have been demonstrated to play an important role in the causal pathway between genetic variants and disease [10].

However, principled methods are needed to explore the relationship between genetic variants, gene expression levels, and disease phenotype. Classical approaches such as genome-wide association studies (GWAS) and eQTL mapping studies identify relationships between variants and outcomes independently, without considering the causal chain of events such as a variant affecting expression which influences phenotype [16, 37]. These approaches consider only pairwise associations, and we are unable to use isolated association studies to elucidate the molecular mechanisms by which multiple genetic variants contribute to phenotype [38, 17]. In order to learn biological mechanisms of disease, integrative analyses of different types of genetic and genomic data is of increasingly significant importance and must improve to represent context-specific biological relationships and be reproducible across studies.

I propose two new methods of integrative analysis of different types of data to identify genetic variants that influence cellular processes to alter phenotype by (1) using the causal inference framework to test for genetic variants whose disease phenotypic effect is mediated through intermediate molecular phenotypes such as gene expression, and (2) use network analysis to identify groups of genetic variants and genes that work together to collectively drive disease phenotypes.

Mediation provides a framework for identifying the cellular mechanisms that drive phenotype, where genetic variation is suggested to be mediated through intermediate variables [14, 13]. This approach proceeds to decompose the total effect of an exposure on an outcome into (1) the Natural Direct Effect (NDE) of the exposure on the outcome and (2) the Natural Indirect Effect (NIE) of the exposure on the outcome through the mediator as demonstrated in Figure 1 [24, 29, 34, 21, 28]. The NDE measures the effect of the exposure on the outcome through other biological pathways. The NIE measures the effect of the exposure on the outcome through the mediator. The structure imposed by mediation has proved important in explaining biological pathways in lung cancer. For instance, three GWAS studies performing simple association analyses identified variants at the 15q25.1 locus associated with increased risk of lung cancer. However, it was unclear whether this variant contributed directly to cancer pathogenesis or indirectly through increased smoking behavior through nicotine dependence [6]. Vanderweele et al used mediation analysis as a statistical framework to directly assess the causal pathway of the 15q25.1 locus. They showed that the effect of candidate genetic variants on lung cancer do not act primarily through smoking intensity [35]. Thus the impact of genetic variants in influencing lung cancer must be working through other pathways to contribute to tumor genesis. One potential path of effect of a genetic variant on a phenotypic outcome is through genomic variants such as gene expression. Thus the eQTL relationship is captured through the indirect or mediated effect, where the variants act through the gene expression to affect the outcome. Under mediation, we expect loci that are not eQTLs to have no indirect effect and only act through other biological mechanisms in the direct effect [14].

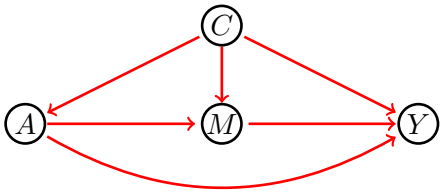

Figure 1: The directed acyclic graph characterizing the mediation framework, relating an exposure A to an outcome Y through mediator M where C is a potential confounder.

To perform a mediation analysis when the phenotypic outcome is binary such as diagnosis of a lung disease, Vanderweele and Vansteelandt derived closed-form results on the odds ratio scale for rare binary outcomes [36, 34]. They provide the identifying assumptions of natural indirect and direct effects in an odds ratio setting. The primary assumption made to obtain closed-form solutions for the effect odds ratios is that the binary outcome is rare, or has a prevalence of below 5%. However, this is not an appropriate assumption for common complex diseases. For instance, we may be interested in the outcome of asthma diagnosis which has prevalence above this rare level in an aging urban population.

Approaches for mediation in the setting of a common dichotomous outcome have been proposed with significant limitations. Tchetgen Tchetgen showed that by assuming that the mediator follows a Bridge distribution one can obtain a closed-form solution for the natural direct and indirect effects [32]. However, this requires the imposition of an additional distribution assumption that oftentimes may not be appropriate for the data. Imai, Keele and Tingley have considered using probit models to directly model the binary outcome for mediation analysis [15]. However, this does not have a clear interpretation. **In Aim 1, I propose a method for mediation under the setting of a common binary outcome without imposing additional assumptions.** The estimation procedure proceeds using the standard regression approach of mediation but uses a probit approximation to obtain closed-form estimators. Variances are also derived to allow for testing of the effects. We will demonstrate this approach in simulation with parameters approximating those one would obtain in an analysis of gene expression mediating genetic variants.

Network analyses have also emerged as an integrative approach to characterize complex genomic associations [3]. Bipartite networks are a natural representation for eQTL associations as shown in Figure 2, where the edges between SNPs and gene expression represent the strength of the eQTL association [2, 25, 4]. Features of a network can inform function. For example, nodes that are more densely connected can represent natural divisions of functional relatedness, in that genetic variants and genes group into highly modular communities. This representation has been shown to identify biological effects in chronic obstructive pulmonary disease (COPD) [11]. In COPD, GWAS-identified single nucleotide polymorphisms (SNPs) were found to be most central among groups of functionally related features [25, 8, 23].

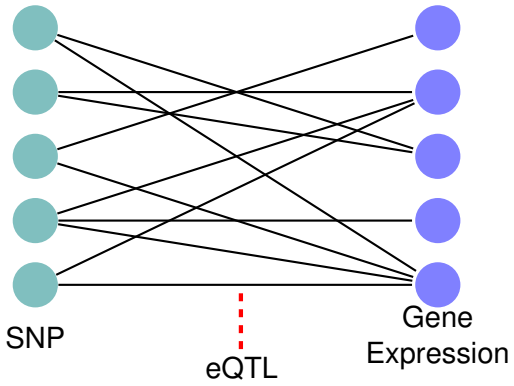

Figure 2: The bipartite network connecting SNP and gene expression.

These methods treat edges (such as the strength of eQTL associations) as fixed and known when these associations are in fact estimated in the initial eQTL analysis [1]. In this initial analysis, SNP and gene expression measurements that have been estimated from an array or sequencing are used in the simple regression model to assess eQTLs. Regression models are relatively robust to minor errors so the measure of association should be statistically consistent. Nonetheless, the associations are estimated, they thus have uncertainty and error. Ignoring the variability of the estimated associations can be detrimental to ensuring results are true and reproducible.

Methods that account for variation within the eQTL network model, particularly due to the error propagated from the preliminary eQTL analysis, are limited and detract from potential reproducibility. Permutation tests are used to assess the confidence associated with each edge in the network [31]. Otherwise, metrics assessing features of a network are reported under the assumption that the edges, or eQTL associations, are fixed and the network has been correctly identified. **In Aim 2, I address this limitation by proposing methodology for incorporating error into estimated network metrics that inform biological relationships, such as degree centrality, which is the number of edges connected to a node.** The degree is a measure of centrality that is associated with how essential a node is to function. We propose both to demonstrate our approach using eQTLs obtained from a cohort study of chronic obstructive pulmonary disease (COPD) and to provide scientific interpretation having propagated error estimates through the network.

B Approach

B.1 AIM 1: Extend mediation methodology for integrative genomic analysis with common binary outcomes between eQTL and network analyses

The goal of Aim 1 is to develop methods for mediation analysis to characterize a mediation relationship with a binary outcome that is common. This is oftentimes the case when we have genetic variants as the exposure, gene expression as the mediator and a common phenotypic or disease outcome. In this setting, we define A an exposure of interest, M the mediator, C a set of baseline confounders and Y the binary, common outcome. We will consider both the settings where M is a continuous mediator and a dichotomous mediator, resulting in different estimators.

B.1.1 Background

Following the traditional approach to mediation, I first make the assumption that M is continuous and normally distributed. For this setting, the outcome Y is binary and common. Thus, we fit the following two regression models for Y and M:

$$logit(P(Y=1|a,m,c)) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta_4' c$$
$$E[M|a,c] = \beta_0 + \beta_1 a + \beta_2' c$$

The causal effects of interest are the natural indirect effect and natural direct effect. Counterfactual notation from the causal inference literature is used, where Y_a and M_a are the the outcome and mediator values when the exposure is set to a. The natural direct effect (NDE) on the odds ratio scale has been defined and compares the odds of the outcome with exposure a and mediator M_{a^*} to the odds of the outcome with exposure a^* and mediator still M_{a^*}, conditional on any covariates C. It is thus a measure of the effect on the outcome through mechanisms other than the mediator. The natural indirect effect (NIE) is a measure of the effect on the outcome through the mediator, thus compares the odds of the outcome with exposure a and mediator M_a to the odds of the outcome with exposure a and mediator M_{a^*} [36, 24, 29, 34].

B.1.2 Proposed Method

The natural indirect effect odds ratio (OR^{NIE}) can be found using the definition of the effect in the log transformed version. We can proceed to use the counterfactual definitions of the components of this term in order to identify the integrals of interest as is the approach standardly taken [33, 19].

$$\begin{aligned}
log(OR^{NIE}_{a,a^*|c}(a)) &= log\left(\frac{P(Y_{aM_a}=1|c)/(1-P(Y_{aM_a}=1|c))}{P(Y_{aM_{a^*}}=1|c)/(1-P(Y_{aM_{a^*}}=1|c))}\right) \\
&= logit(P(Y_{aM_a}=1|c)) - logit(P(Y_{aM_{a^*}}=1|c)) \\
&= \int P(Y=1|a,c,m)f(m|a,c)dm - \int P(Y=1|a,c,m)f(m|a^*,c)dm \\
&= \int expit(\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c)f(m|a,c)dm \\
&- \int expit(\theta_0+\theta_1 a+\theta_2 m+\theta_3 am+\theta_4' c)f(m|a^*,c)dm
\end{aligned}$$

This integral does not have a closed-form solution; however, the relationship between the logit and probit models can be exploited. Particularly, $\frac{e^a}{l+e^a} \approx \Phi(sa)$ where s is defined such that the variances of the distributions are scaled equally. Extensive simulation has identified that multiplying logit coefficients by $s = 1/1.6$ to be the nearest approximation to probit coefficients. The distributions are each symmetric about 0.5 and diverge in the tails, making the approximation appropriate when the proportion with the outcome is between $0.2-0.7$. I consider the inverse of this approximation to the logit model to obtain a closed form solution of the NIE OR. Defining the odds function $H(p) = p/(1-p)$,

$$OR^{NIE}_{a,a^*|c}(a) \approx \frac{H\left(\Phi\left(\frac{s\theta_0+s\theta_1 a+s\theta_4' c+(s\theta_2+s\theta_3 a)(\beta_0+\beta_1 a+\beta_2' c)}{\sqrt{1+(s\theta_3 a\sigma+s\theta_2\sigma)^2}}\right)\right)}{H\left(\Phi\left(\frac{s\theta_0+s\theta_1 a+s\theta_4' c+(s\theta_2+s\theta_3 a)(\beta_0+\beta_1 a^*+\beta_2' c)}{\sqrt{1+(s\theta_3 a\sigma+s\theta_2\sigma)^2}}\right)\right)}$$

The standard errors of the OR^{NIE} are critical for characterizing the uncertainty in the estimates. The errors can be approximated using the multivariate delta method, which utilizes a first-order Taylor approximation to estimate errors of functions of random variables. This is given by $SE(OR^{NIE}) \approx \sqrt{\Gamma\Sigma\Gamma'}|a-a^*|$ where the components of this formula are defined to be $\Sigma \equiv \begin{pmatrix} \Sigma_\beta & 0 & 0 \\ 0 & \Sigma_\theta & 0 \\ 0 & 0 & \Sigma_{\sigma^2} \end{pmatrix}$ and $\Gamma = (d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8)$. Note that Σ_β, Σ_θ and Σ_{σ^2} are the covariance matrices for the estimators of $\beta, \mathbf{\Theta}, \sigma^2$. Additionally, the values of Γ are the partial derivatives of OR^{NIE}. The delta method has been demonstrated to perform poorly in many mediation settings [20]. Other methods to assess the error of the effects include taking bootstrap samples and computing

the standard error or the median absolute deviation. I will assess each of these ways for calculating the error in order to recommend a best approach in this setting. The natural direct effect odds ratio (OR^{NDE}) can be found by again using the effect definition and leveraging the relationship between the logit and probit distributions,

$$OR^{NDE}_{a,a^*|c}(a^*) = \frac{P(Y_{aM_{a^*}} = 1|c)/(1 - P(Y_{aM_{a^*}} = 1|c))}{P(Y_{a^*M_{a^*}} = 1|c)/(1 - P(Y_{a^*M_{a^*}} = 1|c))} = \frac{H\left(\Phi\left(\frac{s\theta_0 + s\theta_1 a + s\theta_4' c + (s\theta_2 + s\theta_3 a)(\beta_0 + \beta_1 a^* + \beta_2' c)}{\sqrt{1+(s\theta_3 a\sigma + s\theta_2\sigma)^2}}\right)\right)}{H\left(\Phi\left(\frac{s\theta_0 + s\theta_1 a^* + s\theta_4' c + (s\theta_2 + s\theta_3 a^*)(\beta_0 + \beta_1 a^* + \beta_2' c)}{\sqrt{1+(s\theta_3 a^*\sigma + s\theta_2\sigma)^2}}\right)\right)}$$

The standard errors of the OR^{NDE} can be found in the same way as the OR^{NIE}, simply updating the values of Γ to reflect this new estimator. Estimators for a binary mediator can be found using the same approach.

B.1.3 Preliminary Results

I conducted a simulation study with parameters set to numerically represent a mediation model with a SNP as the exposure, gene expression at the same locus as the mediator and a common binary disease status. The exposure variable A (SNP) was simulated from a Bernoulli distribution with $p = 0.4$. A continuous covariate C was simulated from $N(0.3A, 1)$. The mediator M (gene expression) and outcome Y (disease status) are simulated as

$$m \sim Norm(0.5a + 0.5c, 1)$$
$$y \sim Bernoulli(expit(k + 0.5a + 0.5m + 0.25c))$$

I varied k from $[-2.75, 0.75]$ at 0.005 intervals in order to specify a disease status with a common prevalence level. The OR^{NIE}, OR^{NDE} and measures of their standard error were averaged across 1,000 simulations for each value of k. The effects were obtained using the proposed method as well as via numerical analysis in order to assess the performance of the approximation and determine where it is appropriate. The standard error was evaluated empirically for reference, via the delta method to assess this closed form of the error, and by bootstrapping using 500 samples.

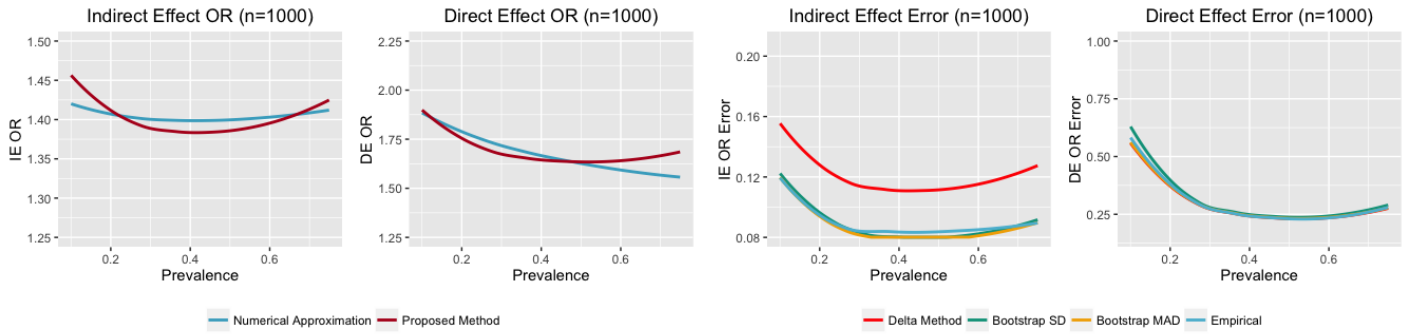

Figure 3. Effect and error estimates of the mediation simulation study.

This simulation study demonstrated that the proposed method is appropriate generally in the same region in which the logit-probit approximation is close, or between 20% and 60% prevalence then diverges in the tails. The error estimates, with respect to their closeness to the empirical error, were all close for the OR^{NDE} in the rightmost panel of Figure 3. However, the delta method overestimated the error in the OR^{NIE} thus the bootstrap approach appears more appropriate in this setting.

B.1.4 Expected outcomes, limitations, alternative approaches

I have calculated closed form estimators for mediation with a common binary outcome that can be computed efficiently. Preliminary results suggest that the median absolute deviation of bootstrap samples is an appropriate estimator for their error. This method has been programmed in R and will be finalized as an R package with example data and vignettes for distribution. Alternative approaches include using a computationally optimized numerical approximation towards estimating the effect. This could be done using Gaussian quadrature or other approximation methods.

I will apply this method to data from the Normative Aging Study (NAS) to assess whether the effect of smoking on asthma is mediated through DNA methylation, a genomic feature that like gene expression has been pro-

posed to be an intermediate of the relationship between genotype and disease [26, 18, 5]. NAS is a prospective cohort study on an aging population in the greater Boston area. The study has data on $n = 1455$ individuals, when reduced to include non-duplicated observations with both clinical and genomic data. The clinical features including smoking history were assessed during study visits at which time blood samples were also collected to assess genomic features; methylation was measured in blood using Illumina Infinium HumanMethlation450 Beadchips. Within this aging population, asthma is a common binary outcome and is thought to be influenced by smoking. I will use the proposed mediation method to test whether smoking acts through methylation or through other biological mechanisms to increase asthma risk.

I will further explore the inference methods that perform best in this mediation setting. Although the primary focus of my work will be to derive estimators and error estimates of the mediation effects, I also plan to determine which effects are statistically significant to inform biological pathways. The most widely used test to determine the significance of the indirect effect, the Sobel test, generally lacks power to detect significant effects [20]. I will explore the use of the tests such as the Berk-Jones statistic that take advantage of the correlation between genomic measures of nearby loci in order to boost weak, sparse signals.

B.2 AIM 2: Develop uncertainty measure for error propagation for integrative genomic analysis via bipartite networks

The overall goal of Aim 2 is to incorporate error measures into network metrics estimated from bipartite eQTL networks to ensure reproducibility and hypotheses that are more likely to be clinically valuable. In Aim 1, I use the mediation approach to obtain effect and error estimates of complex relationships of different genomic data types. However, mediation analysis is best suited to testing a handful of genetic variants that are likely to influence traits. When traits depend on hundreds or thousands of variants, network methods are well suited to handle this complexity and make fewer assumptions about the nature of the associations between SNPs and genes.

eQTLs are identified by assessing the association between two data types, SNP genotypes (S) and gene expression (G) [16, 37]. Those associations are then cast as edges between SNPs and genes in a bipartite network, which is represented by an adjacency matrix. The adjacency matrix consists of elements $\{i, j\}$, which represent an association between nodes i and j. The adjacency matrix for a bipartite eQTL network is given as $\mathbf{A} = \left(\begin{smallmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{0} \end{smallmatrix}\right)$ where $\mathbf{0}$ is an $n \times n$ matrix of zeroes and $\mathbf{B}$ is an $S \times G$ matrix indicating the relationship between elements of S and G. The elements $A_{i,j}$ of the adjacency matrix outside of $\mathbf{0}$ can be defined (1) continuously using the eQTL measures of association β directly as elements of $\mathbf{B}$ or (2) by dichotomizing all potential eQTL relationships according to a cutoff, such as the p-value for the association test or based on the false discovery rate.

We can develop error estimates for the network metrics by incorporating the error from the eQTL associations in the adjacency matrix. While many network metrics are used, the degree centrality is an important measure and often approximates more complicated measures well and thus will be the metric of interest. The degree of node i is defined to be $C_i^D = \sum_{j,j\neq i} y_{i,j}$, where $y_{i,j}$ is an indicator equal to 1 when nodes i and j are connected by an edge, meaning that within $A_{i,j}$ their shared element is non-zero, and otherwise $y_{i,j} = 0$. The degree of a node is therefore simply the number of edges connected to that node. For an eQTL network, this identifies SNPs that are most highly linked to the expression of genes and therefore should be highly functionally relevant. By estimating the error in this measure, we can test whether GWAS SNPs are not only central to COPD networks, but have a reproducibly high degree [25, 8, 23]. Further, we can identify which SNPs regulate more genes and acquire functional relevance in a robust and disease-specific manner. This mathematically translates to finding the nodes with a statistically significant increase in degree between different disease studies.

B.2.1 Proposed method

I propose an approach to modeling the the error of degree as a function of eQTL associations.

Continuous adjacency matrix: The elements of A are equal to $\beta_{i,j}$ from the corresponding eQTL regression. Thus the variance of the degree is also directly with respect to the $\beta_{i,j}$:

$$\begin{aligned} Var(C_i^D) &= Var\left(\sum_{j,j\neq i} y_{i,j}\right) = Var\left(\sum_{j,j\neq i} I(\beta_{i,j} \neq 0)\right) \\ &= \sum_{j,j\neq i} Var(I(\beta_{i,j} \neq 0)) + \sum_{j,k,j\neq i,k\neq i} Cov(I(\beta_{i,j} \neq 0), I(\beta_{i,k} \neq 0)) \end{aligned}$$

Dichotomous adjacency matrix: The elements of A are equal to an indicator of a particular cut-off. We will follow the same FDR cut-off of 0.1 considered in Platig 2016, so that elements correspond to $I(FDR(\beta_{i,j}) < 0.1)$ from the corresponding eQTL regression [25]. Thus the standard error of the degree is not with direct respect to the $\beta_{i,j}$, rather directly related to the FDR:

$$\begin{aligned} Var(C_i^D) &= Var\left(\sum_{j,j\neq i} y_{i,j}\right) = Var\left(\sum_{j,j\neq i} I(FDR(\beta_{i,j}) < 0.1)\right) \\ &= \sum_{j,j\neq i} Var(I(FDR(\beta_{i,j}) < 0.1)) + \sum_{j,k,j\neq i,k\neq i} Cov(I(FDR(\beta_{i,j}) < 0.1), I(FDR(\beta_{i,k}) < 0.1)) \end{aligned}$$

We can use the formulas from the eQTL regression analysis and its corresponding false discovery rate to assess the first variance summation in each formula, $\sum_{j,j\neq i} Var(I(\beta_{i,j} \neq 0))$ or $\sum_{j,j\neq i} Var(I(FDR(\beta_{i,j}) < 0.1))$. To estimate the covariance between loci, we will use the linkage disequilibrium estimated from the data used in the eQTL association. Comparison can then be given to whether continuous or dichotomous matrices are more appropriate based on both scientific interpretation and error of the scientific metrics. Given this measure of error, a statistical test can be developed to assess between networks whether a node differs in degree or is not significantly different in how connected it is. This is important for the validation of findings, as one could statistically test for consistency of network metrics across networks to demonstrate reproducibility. It would also allow for the comparison of networks to assess if there is consistency in genes of regulatory impact across diseases. The statistical test would utilize the previously defined point estimate of the degree as well as the derived measures of error.

B.2.2 Expected outcomes, limitations, alternative approaches

The outcome of this work will be formulas of error estimates and their software implementation for a bipartite network representing eQTL links. In order to ensure that this method is reasonably implementable, the computational steps will need to be parallelized and perhaps treated as sparse, given that eQTL analysis is traditionally performed across all pairwise relationships between SNP and gene expression.

To demonstrate the impact on the reproducibility and interpretation of networks, I will work with Drs. Quackenbush and Silverman to apply this method to a study of COPD. We will use data from the COPDGene study [27]. COPDGene is a prospective cohort study that was conducted across multiple centers with over 10,000 current and former smokers. The study collected clinical information, SNP genotyping, CT scans and RNA-seq gene expression data for $n = 525$ subjects. All computation will be performed on a high performance cluster using parallelized and multithreaded approaches in R. The data will be prepared by matching genotypes and expression, removing samples with over 10% missingness and minor allele frequency of SNPs less than 0.05. The method will be assessed by comparing the constructed network to an eQTL network previously identified for COPD using data from the Lung Genomics Research Consortium [25].

Accounting for the variance in the intermediate eQTL analysis can also be done within the procedure of network identification. Specifically, a representation of the error can be incorporated using weights on the edges between nodes. The coefficient of determination, R^2, from the eQTL analysis could be used as a weight. The coefficient of determination here is the proportion of variance in gene expression explained by the SNP. This would allow for error to be incorporated in the network; however, would not allow for clear statistical inference approaches.

C Summary

In Aim 1, I develop mediation methods for the setting of a common binary outcome with applications to genomic data. In Aim 2, I investigate integrative genomic analyses using network approaches while accounting for propagated error. These methods fill gaps in the literature of both mediation and network methodology, respectively. These methods will be applied to asthma and COPD applications with the goal of elucidating the complex mechanisms of these diseases acting through genomics. All methods will be programmed and made available in R packages to encourage others to use towards integrative genomic analysis. These aims will give me experience in working with multiple genomic data types, computing, collaborating with experts, and developing statistical methods that address the needs of interdisciplinary scientific questions.

References

[1] Frank W Albert and Leonid Kruglyak. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, 16(4):197–212, 2015.

[2] Armen S Asratian, Tristan MJ Denley, and Roland Häggkvist. *Bipartite graphs and their applications*, volume 131. Cambridge University Press, 1998.

[3] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[4] Michael J Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, 2007.

[5] Benjamin Bell, Charles L Rose, and Albert Damon. The normative aging study: an interdisciplinary and longitudinal study of health and aging. *The International Journal of Aging and Human Development*, 3(1):5–17, 1972.

[6] Stephen J Chanock and David J Hunter. Genomics: when the smoke clears... *Nature*, 452(7187):537–538, 2008.

[7] Vivian G Cheung, Richard S Spielman, Kathryn G Ewens, Teresa M Weber, Michael Morley, and Joshua T Burdick. Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437(7063):1365–1369, 2005.

[8] Michael H Cho, Merry-Lynn N McDonald, Xiaobo Zhou, Manuel Mattheisen, Peter J Castaldi, Craig P Hersh, Dawn L DeMeo, Jody S Sylvia, John Ziniti, Nan M Laird, et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The lancet Respiratory medicine*, 2(3):214–225, 2014.

[9] GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.

[10] Emmanouil T Dermitzakis. From gene expression to disease risk. *Nature genetics*, 40(5):492–493, 2008.

[11] Kimberly Glass, John Quackenbush, Edwin K Silverman, Bartolome Celli, Stephen I Rennard, Guo-Cheng Yuan, and Dawn L DeMeo. Sexually-dimorphic targeting of functionally-related genes in copd. *BMC systems biology*, 8(1):118, 2014.

[12] R David Hawkins, Gary C Hon, and Bing Ren. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11(7):476–486, 2010.

[13] Yen-Tsung Huang, Liming Liang, Miriam F Moffatt, William OCM Cookson, and Xihong Lin. igwas: Integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. *Genetic epidemiology*, 39(5):347–356, 2015.

[14] Yen-Tsung Huang, Tyler J VanderWeele, and Xihong Lin. Joint analysis of snp and gene expression data in genetic association studies of complex diseases. *The annals of applied statistics*, 8(1):352, 2014.

[15] Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, pages 51–71, 2010.

[16] CM Kendziorski, M Chen, M Yuan, H Lan, and AD Attie. Statistical methods for expression quantitative trait loci (eqtl) mapping. *Biometrics*, 62(1):19–27, 2006.

[17] Arthur Korte and Ashley Farlow. The advantages and limitations of trait analysis with gwas: a review. *Plant methods*, 9(1):1, 2013.

[18] Johanna Lepeule, Andrea Baccarelli, Letizia Tarantini, Valeria Motta, Laura Cantone, Augusto A Litonjua, David Sparrow, Pantel S Vokonas, and Joel Schwartz. Gene promoter methylation is associated with lung function in the elderly: the normative aging study. *Epigenetics*, 7(3):261–269, 2012.

[19] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation analysis. *Annual review of psychology*, 58:593, 2007.

[20] David P MacKinnon, Chondra M Lockwood, Jeanne M Hoffman, Stephen G West, and Virgil Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, 7(1):83, 2002.

[21] David Peter MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2008.

[22] Michael Morley, Cliona M Molony, Teresa M Weber, James L Devlin, Kathryn G Ewens, Richard S Spielman, and Vivian G Cheung. Genetic analysis of genome-wide variation in human gene expression. *Nature*, 430(7001):743–747, 2004.

[23] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS Genetics*, 6(4):e1000888, 2010.

[24] Judea Pearl. Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc., 2001.

[25] John Platig, Peter Castaldi, Dawn DeMeo, and John Quackenbush. Bipartite community structure of eqtls. *PLoS Computational Biology*, 12(9):e1005033, 2016.

[26] Weiliang Qiu, Andrea Baccarelli, Vincent J Carey, Nadia Boutaoui, Helene Bacherman, Barbara Klanderman, Stephen Rennard, Alvar Agusti, Wayne Anderson, David A Lomas, et al. Variable dna methylation is associated with chronic obstructive pulmonary disease and lung function. *American journal of respiratory and critical care medicine*, 185(4):373–381, 2012.

[27] Elizabeth A Regan, John E Hokanson, James R Murphy, Barry Make, David A Lynch, Terri H Beaty, Douglas Curran-Everett, Edwin K Silverman, and James D Crapo. Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2011.

[28] James M Robins. Semantics of causal dag models and the identification of direct and indirect effects. *Highly struc-tured stochastic systems*, pages 70–81, 2003.

[29] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

[30] Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005.

[31] Sean L Simpson, Robert G Lyday, Satoru Hayasaka, Anthony P Marsh, and Paul J Laurienti. A permutation testing framework to compare groups of brain networks. *Frontiers in computational neuroscience*, 7:171, 2013.

[32] Eric Tchetgen Tchetgen. A note on formulae for causal mediation analysis in an odds ratio context. *Epidemiologic methods*, 2(1):21–31, 2014.

[33] Linda Valeri and Tyler J VanderWeele. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with sas and spss macros. *Psychological methods*, 18(2):137, 2013.

[34] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[35] Tyler J VanderWeele, Kofi Asomaning, Eric J Tchetgen Tchetgen, Younghun Han, Margaret R Spitz, Sanjay Shete, Xifeng Wu, Valerie Gaborieau, Ying Wang, John McLaughlin, et al. Genetic variants on 15q25. 1, smoking, and lung cancer: an assessment of mediation and interaction. *American journal of epidemiology*, 175(10):1013–1020, 2012.

[36] Tyler J VanderWeele and Stijn Vansteelandt. Odds ratios for mediation analysis for a dichotomous outcome. *American journal of epidemiology*, 172(12):1339–1348, 2010.

[37] Kai Wang, Mingyao Li, and Hakon Hakonarson. Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, 11(12):843–854, 2010.

[38] Lucas D Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–1106, 2012.